The Hesitating Robot - Implementation and First Impressions

Birte Carlmeyer*†‡
bcarlmey@techfak.uni-bielefeld.de

Simon Betz*‡§
simon.betz@uni-bielefeld.de

Petra Wagner*§
petra.wagner@uni-bielefeld.de

David Schlangen*‡
david.schlangen@uni-bielefeld.de

Britta Wrede*†
bwrede@techfak.uni-bielefeld.de

ABSTRACT

In this paper we present the implementation of a robot, that dynamically hesitates, based on the attention of the human interaction partner. To this end, we outline requirements for a real-time interaction scenario, describe the realization of a disfluency insertion strategy, and present observations from the first tests of the system.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; • **Computing methodologies → Natural language processing**;

ACM Reference Format:
Birte Carlmeyer, Simon Betz, Petra Wagner, David Schlangen, and Britta Wrede. 2018. The Hesitating Robot - Implementation and First Impressions. In *HRI '18 Companion: 2018 ACM/IEEE International Conference on Human-Robot Interaction Companion, March 5–8, 2018, Chicago, IL, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3173386.3176992

1 INTRODUCTION

Humans observe their interaction partner, among other reasons, in order to detect communication mistakes or figure out how attentive the interlocutor is in the conversation. Based on feedback signals we adapt our behavior to manage conversations. One adaptation strategy is hesitation, through pauses, repetitions, or fillers, to obtain extra time without losing the conversational floor. Additionally, [8] indicate that hesitations heighten the listeners' attention to upcoming speech. To have a situated human-robot interaction (HRI), it is necessary that the robot can recognize human feedback signals (e.g., social eye-gaze [1]) and is able to react to them, (e.g., with hesitations).

Little analysis has been conducted on the impact of hesitations on human-robot interaction. [4] use filled pauses ("So..." "Let's see...") between dialogue acts to give the system more time for perception and decision making. The authors evaluated the disengagement costs and could show, that the use of hesitations in combination with a forecasting model of disengagement lead to less costly disengagement management, but no evaluation of the effect on the human interaction partner, e.g., on task performance or subjective rating of the robot are conducted.

[6, 7] evaluated self-interruptions in an smart-home setting. The authors use silence as an attention-regaining strategy whenever the attention of the human interaction partner moves away. They could show, that this strategy has a positive effect on the visual attention of the human, but at the cost of less positive subjective ratings. The attention-regaining strategy was effective but the agent was perceived as rude and less friendly.

Based on a human-human interaction corpus analysis, [3] propose a disfluency insertion strategy for synthetic speech, which consist of the following cascade: *(i) lengthening:* add lengthening at the next appropriate syllable *(ii) silence:* insert a silence for a maximum of 1000ms *(iii) filler:* insert a filler (e.g., "uhm") *(iv) silence:* insert an additional silence. Whereas in [7] the attention-regaining strategy consists of a simple pausing of the synthesis (while the interaction partner is not attentive), this strategy has several levels of escalation. To counteract the perceived rudeness observed in [6], we integrate this hesitation strategy into our dialogue system for smoother transitions from fluent delivery to hesitation mode.

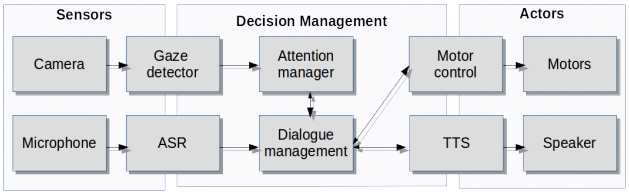

Figure 1: System overview of the main components.

2 REQUIREMENTS & IMPLEMENTATION

To manage the dialogue between a human and the robot, two main questions have to be answered: (i) *When to (re-)act?* and (ii) *How to (re-)act?* To implement hesitations, based on the attention of the interaction partner, several requirements have to be met.

Perception of the interaction partner. The robot needs to be able to perceive the human interaction partner. To answer the first question - *when to react?* - it is necessary to observe the interlocutor. For this purpose we use, among other things, a gaze detector [9] to assess the current visual focus of attention (VFoA).

Model of attention. Based on the perception, it has to decide whether the interaction partner is attentive or not. The robot needs an internal concept of attention. Here we define attention as a state in which the VFoA matches with current focus of discourse (FoD)(cf. [6]).

*Cluster of Excellence Cognitive Interaction Technology (CITEC), Bielefeld University
†Applied Informatics Group, Faculty of Technology, Bielefeld University
‡Dialogue Systems Group, Faculty of Linguistics and Literary Studies
§Phonetics and Phonology Workgroup, Faculty of Linguistics and Literary Studies

HRI '18 Companion, March 5–8, 2018, Chicago, IL, USA

ACM ISBN 978-1-4503-5615-2/18/03.
https://doi.org/10.1145/3173386.3176992

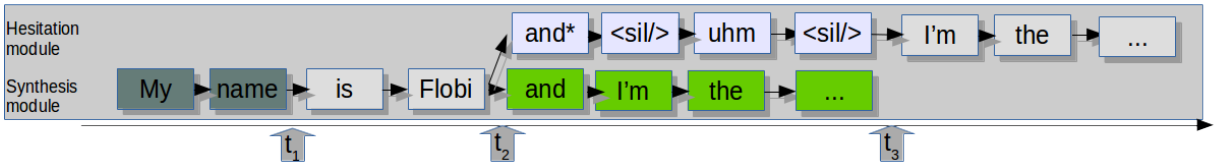

Figure 2: Production of the *wordIU's* of the *hesitation module*: (t_1) module receives *hesitation start* event; (t_2) best entry point for the hesitation; (t_3) module receives *hesitation end* signal. Color scheme of the different incremental units: (dark gray) already synthesized; (blue) hesitation insertions; (green) revoked *wordIU's*.

Incremental processing capabilities. In order to realize the proposed hesitation insertion strategy - in this case the answer of the question *how to react?* - we need the possibility to change the ongoing speech plan. Therefore the capability of incremental processing, especially on the synthesis level, is mandatory. We realize this by using the incremental processing framework *InproTK* [2].

Figure 1 shows an overview of the main parts of the current system. The agent can perceive the human interaction partner via several sensors, in this case via a microphone and a small webcam. The attention management receives information from the gaze detector [9] - the VFoA - and context information from the dialogue management - the current FoD - and provides information about the estimated attention state of the human interaction partner (*attentive*, *not-attentive*). The dialogue management [5] can trigger the motor control of the robot, e.g., to show attention at the current FoD by looking at it and is able to send utterances to text-to-speech synthesis module (tts). Based on the attention state, the dialogue manager can also start and stop the *hesitation strategy*.

Figure 2 shows an example of the hesitation strategy, which is implemented as separate module in *InproTK*, an implementation of the general, abstract model for incremental processing [10]. The model consists of a network of processing modules, which exchange incremental data in form of incremental units (IU). The IU-modules receive information on their *left buffer*, perform some kind of processing on these IU's and provide output on their *right buffer*. At the moment t_1 the module receives the event *'start hesitation'*. The strategy takes the IUs from the *left buffer* of the synthesis module, in this case a list of *wordIUs*, each representing a single word. It searches for the best entry point and lengthens the most appropriate segments. In this example, the synthesis module already played back the first two *wordIUs* "my" and "name". The rest of the current phrase ("is Flobi and I'm the ...") is still in the playback pipeline. According to the proposed strategy, the best entry point for the hesitation strategy is the *wordIU* "and" (t_2), which is then stretched by a factor based on the findings of [3]. Then the synthesis module will be paused up to 1000ms (< *sil*/>). If this is not enough time, the module inserts a filler ("uhm"), also applied with lengthening, followed by a second pause until the dialogue management stops the hesitation strategy (t_3). In the case the dialogue management wants to stop the hesitation strategy earlier (e.g., the estimated attention state of the human interaction partner changed to *attentive*), the strategy can be interrupted at several points: (i) before the entry point t_2: without any effect in the synthesis (ii) before the filler ("uhm"): the lengthening will be produced, but the silence can be interrupted (iii) during or after the filler: the filler will be produced, but the silence is again interruptible.

3 CURRENT STATE & FUTURE WORK

We tested our system in a small pilot study (n=4) in an interaction scenario in a smart apartment to get some insights and first impressions of our system. As a platform we use the simulation of the anthropomorphic robot head Flobi. The system is mostly working as expected. The hesitation strategy starts if the human is not attentive (looking away) and stops when the user refocuses. We observed some issues, which need to be addressed before we can evaluate our system in an interaction study. The insertion of the filler sometimes leads to noise inferences, which need to be eliminated. Additionally, producing fillers such as "uhm" is not a trivial issue, because they are normally not part of the training corpus for speech synthesis voices, and at least for German, they cannot be synthesized out-off-the-box with the tts system at hand without further acoustic modification. We need to investigate if the participants correctly interpret the intention of these fillers as hesitations. After these issues are solved, the next step will be an evaluation study to investigate the effect of this hesitation strategy on the attention, task progress, and the subjective ratings of the agent in order to test if the lengthening and the insertion of fillers can counteract the in [7] perceived rudeness of self-interruptions.

4 ACKNOWLEDGMENTS

This work was funded as part of the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277), Bielefeld University.

REFERENCES

[1] Henny Admoni and Brian Scassellati. 2017. Social Eye Gaze in Human-Robot Interaction: A Review. 6 (03 2017), 25.
[2] Timo Baumann and David Schlangen. 2012. The InproTK 2012 release *(Proc. of the NAACL-HLT Workshop in SDCTD 2012)*. ACL, 29–32.
[3] Simon Betz, Petra Wagner, and Jana VoÃ§e. 2016. Deriving a strategy for synthesizing lengthening disfluencies based on spontaneous conversational speech data. In *Tagungsband der 12. Tagung Phonetik und Phonologie im deutschsprachigen Raum*. LMU, 19–22.
[4] Dan Bohus and Eric Horvitz. 2014. Managing Human-Robot Engagement with Forecasts and... Um... Hesitations. In *Proc. of the 16th ICMI*. ACM, 2–9.
[5] Birte Carlmeyer, David Schlangen, and Britta Wrede. 2014. Towards Closed Feedback Loops in HRI. In *Proceedings of ICMI-MMRWHRI '14*. ACM, 1–6.
[6] Birte Carlmeyer, David Schlangen, and Britta Wrede. 2016. Exploring self-interruptions as a strategy for regaining the attention of distracted users. In *Proc. of - EISE '16*. ACM.
[7] Birte Carlmeyer, David Schlangen, and Britta Wrede. 2016. "Look at Me!": Self-Interruptions as Attention Booster?. In *Proc. of HAI '16*. ACM.
[8] Philip Collard. 2009. *Disfluency and listeners' attention: An investigation of the immediate and lasting effects of hesitations in speech*. Ph.D. Dissertation. University of Edinburgh.
[9] L. Schillingmann and Y. Nagai. 2015. Yet another gaze detector. In *2015 IEEE-RAS 15th Humanoids*. 8–13.
[10] David Schlangen and Gabriel Skantze. 2011. A General, Abstract Model of Incremental Dialogue Processing. *Dialogue and Discourse* 2, 1 (2011), 83–111.